RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

GÖTEBORGS UNIVERSITET

SWE-CLARIN

SVALA:

Pseudonymizer service for Swedish L2 essays

Elena Volodina, Samir Ali Mohammed, Arild Matsson

RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

GÖTEBORGS UNIVERSITET

SWE-CLARIN

SVALA:

SVA = SVenska som Andraspråk, Linking & Annotation

Pseudonymizer service for Swedish L2 essays

Elena Volodina, Samir Ali Mohammed, Arild Matsson

RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

GÖTEBORGS UNIVERSITET

SWE-CLARIN

SweLL –

research infrastructure for Swedish as a Second Language

RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

GÖTEBORGS UNIVERSITET

SWE-CLARIN

Swedish Learner Language

SweLL –

research infrastructure for Swedish as a Second Language

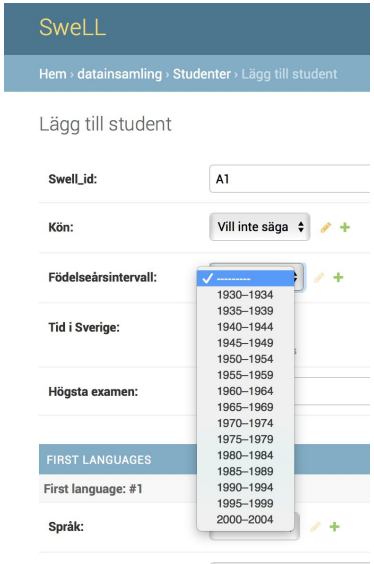
SweLL promises (main)

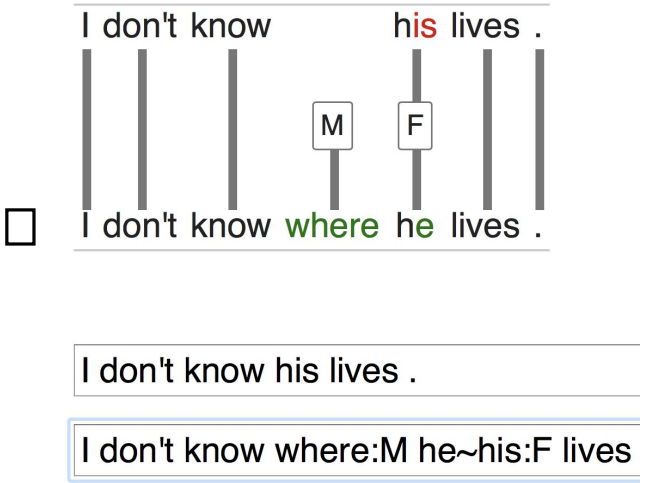
1. Deliver a well-annotated (gold standard) corpus of L2 essays
   - 600 essays, approx 100 per CEFR levels A1-C1 + 100 for control L1 learner corpus
   - Incl manual correction annotation & manually checked linguistic annotation
   - Make available for research (and public?)

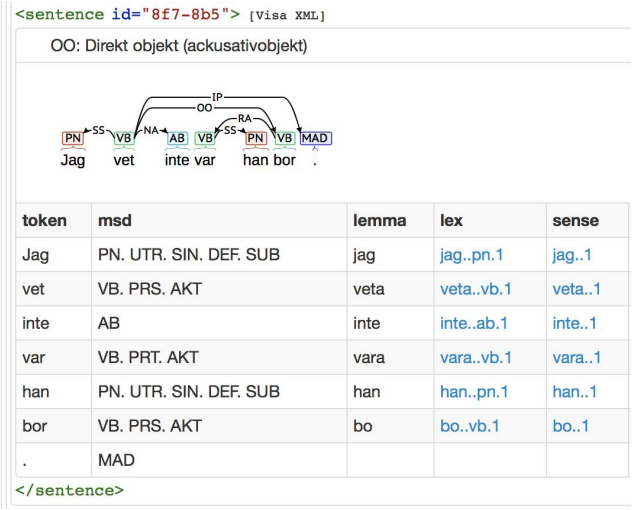
```
graph = {
  "source": [
    {"id": "s0", "text": "I "},
    {"id": "s1", "text": "don't "},
    {"id": "s2", "text": "know "},
    {"id": "s3", "text": "his "},
    {"id": "s4", "text": "lives "},
    {"id": "s5", "text": ". "}
  ],
  "target": [
    {"id": "t0", "text": "I "},
    {"id": "t1", "text": "don't "},
    {"id": "t2", "text": "know "},
    {"id": "t3", "text": "where "},
    {"id": "t4", "text": "he "},
    {"id": "t5", "text": "lives "},
    {"id": "t6", "text": ". "}
  ],
  "edges": {
    "e-s0-t0": {"id": "e-s0-t0", "ids":
    "e-s1-t1": {"id": "e-s1-t1", "ids":
```

SweLL promises (main)

2. Set a platform (and workflow) for
   - Continuous upload of new essays
   - Manual correction annotation
   - Automatic linguistic annotation





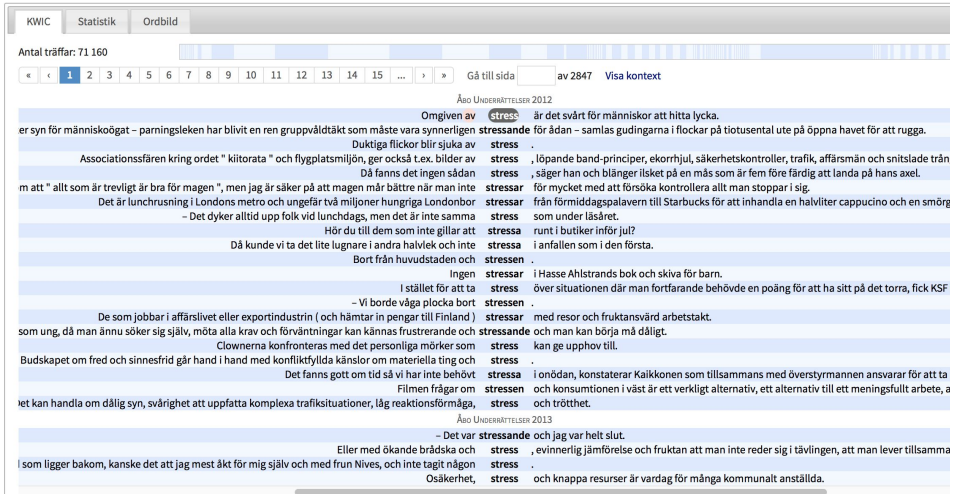


token	msd	lemma	lex	sense
Jag	PN. UTR. SIN. DEF. SUB	jag	jag..pn.1	jag..1
vet	VB. PRS. AKT	veta	veta..vb.1	veta..1
inte	AB	inte	inte..ab.1	inte..1
var	VB. PRT. AKT	vara	vara..vb.1	vara..1
han	PN. UTR. SIN. DEF. SUB	han	han..pn.1	han..1
bor	VB. PRS. AKT	bo	bo..vb.1	bo..1
.	MAD			

</sentence>

SweLL promises (main)

- Set a platform for browsing L2 essays
  - in concordance fashion (+parallel view)
  - In full text fashion



SW1203-UPPSATSER

I Sverige lever många människor det goda livet tycker jag. Det är inte så i många andra land. Människor bryr sig om att äta, träna och sova ordentligt. De vill också ha ett rikt socialt liv vilket är viktigt för psykosocial hälsa. Att ha ett gott liv är något viktigt för mig. Det är ett ständigt jobb. Man måste alltid tänka på sin hälsa. Om man inte har några problem med hälsa måste man träna. Idag sitter vi mycket mer än förrut. Sittande arbete gör oss lat och vi har inte inspiration att börja röra på oss. En familj som ofta vandrar i fjällen, cyklar, promenerar eller lekar ute tillsammans har det goda livet enligt mig. Mat är en av viktigaste saker angoende det goda livet. Man måste välja väldigt noga sin mat. Det är lätt att vara nöjd med halvferdig mat vilket man lagga snabbt. Att lagga riktig nytig mat tar mycket tid och man bör förbereda sig. Jag menar att jag måste köpa färska grönsaker om jag ska lagga någon nytig mat. Jag menar med mat försöker man att äta hälsosamt och undvika fetma, diabetes, hög blodtryck och hjärt och kärlsjukdomar. Det är bäst om man är vegetarisk och icke-rökare. Jag tror att frasen "Det goda livet " ska referera till glad familjen som lever hälsosamt liv utan stress. Å andra sida är jag inte säkert att det är möjligt i ett modernt samhälle leva detta liv. I dagens samhället ar viktigt att tjäna mycket pengar därför att pengarna betyder en hög status och vi alla vill ha hög status.

I modernt samhälle kommer tyvärr stress och många andra negativa saker. Till slutet vill jag säga att det goda livet är mitt mål. Ett foto av lycklig familjen på ett bord.

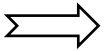
Nu tillbaka till Europa och Sverige. Här har människorna andra problem. Stress, långa

A truly interdisciplinary effort

- SLA
- NLP
- Software engineering

A truly interdisciplinary effort

- SLA
- NLP
- Software engineering

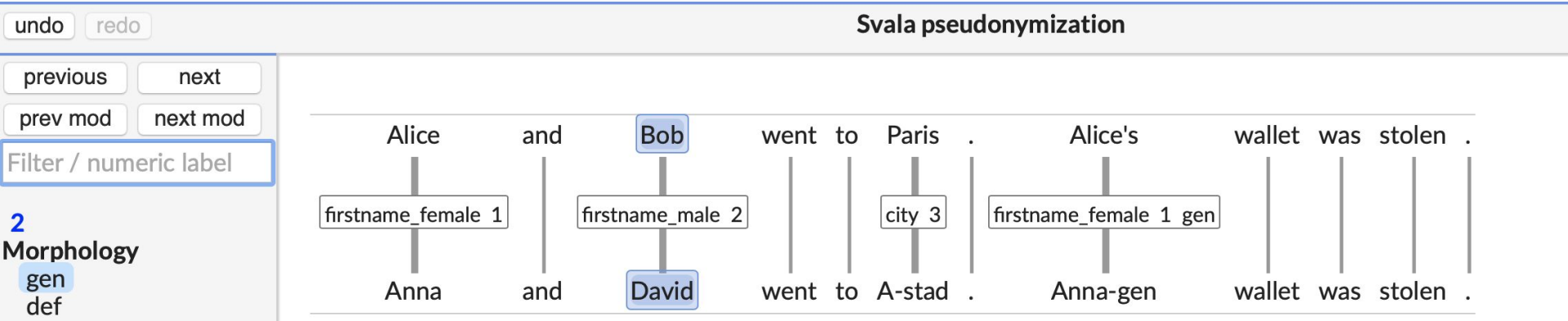


Effects on SVALA

- Methodology of annotation
- Design of SVALA
- Format
- Terminology

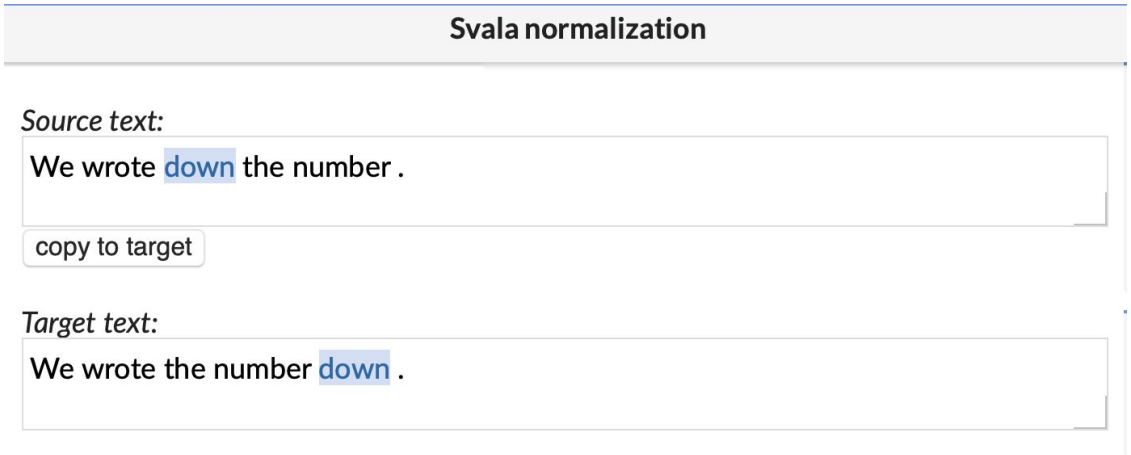
Methodology

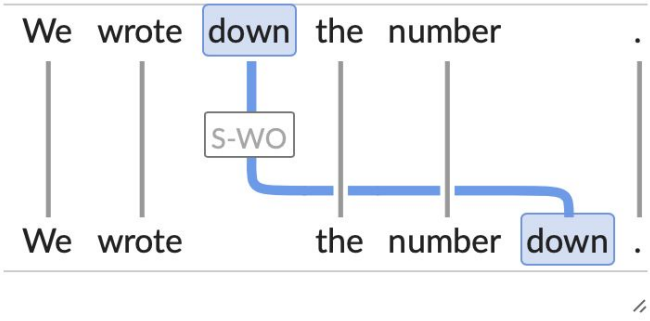
- Anonymization in 3 steps:
  - detection
  - placeholder labelling
  - pseudonymization
- Normalization separate from correction annotation



Design

- As user-friendly (for the target user group) as possible:
  - No xml tags
  - Work in a ”familiar” environment (i.e. text editor)
  - Visual support
  - Support for word-order changes





Format

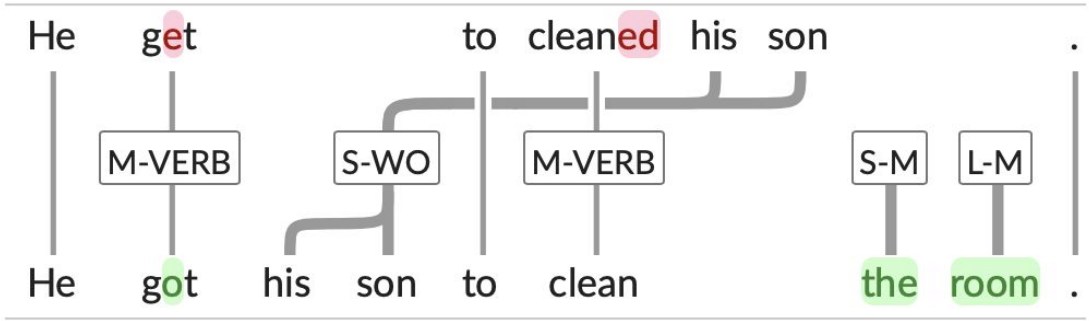
- JSON format
- 3 separate ”objects”:
  - for original text
  - normalized text
  - links with labels


```
{
  "source": [
    {"id": "s0", "text": "We "},
    {"id": "s1", "text": "wrote "},
    {"id": "s2", "text": "down "},
    {"id": "s3", "text": "the "},
    {"id": "s4", "text": "number "},
    {"id": "s5", "text": ". "}
  ],
  "target": [
    {"id": "t0", "text": "We "},
    {"id": "t1", "text": "wrote "},
    {"id": "t2", "text": "the "},
    {"id": "t3", "text": "number "},
    {"id": "t4", "text": "down "},
    {"id": "t5", "text": ". "}
  ],
  "edges": {
    "e-s2-t4": {
      "id": "e-s2-t4",
      "ids": ["s2", "t4"],
      "labels": ["S-WO"],
      "manual": true
    },
```

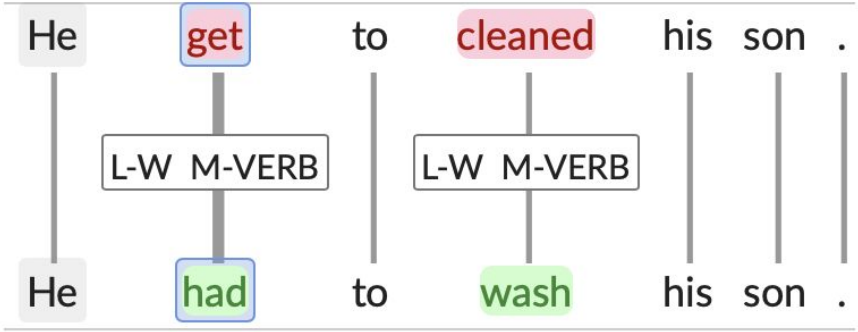
Terminology

- Correction annotation (not error annotation!)

(1)



(2)



SVALA - SweLL annotation tool

- Parallel text
- Visualized diff
- Semi-automatic word alignment
- Annotation on source–target links

Desired:

- Drop-down menus for error codes
- Three-tier representation (original, spell-corrected, normalized)
- Support with automatic spelling error detection

Dan Rosén,
research engineer

Arild Matsson,
research engineer

SVALA - automatic pseudonymizer service

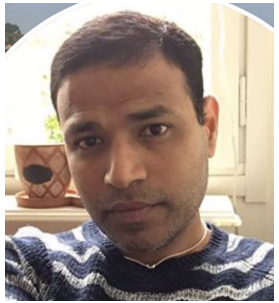
- Why?
  - GDPR
  - Ethical reasons
  - To speed up annotation work
  - To boost essay collection

SVALA - automatic pseudonymizer service

- 8 head categories, 40 subcategories, morph. markers
  - names, geo names, institutions, transportation, age, date, miscellaneous, mark/sensitive

1. ORIGINAL TEXT →@PLACEHOLDER →RENDERING
2. ORIGINAL TEXT →@PLACEHOLDER →REPLACEMENT
3. ORIGINAL TEXT →@PLACEHOLDER →ORIGINAL

Guidelines: https://spraakbanken.github.io/swell-project/Anonymization_guidelines



Samir Ali Mohammed,
Systems developer

Arild Matsson,
research engineer

SVALA - pseudonymizer: facts

- Rule-based
  - tokenization, regular expressions
  - minor spelling correction (Levenstein distance)
- Git Repo: https://github.com/SamirYousuf/LR_project
- Python
- IN: text; OUT: json / text

- Three steps:
  - detection
  - placeholder labelling
  - pseudonymization

SVALA – hands-on demo

https://spraakbanken.gu.se/swell/dev/

Example essay (translation into English + mocking errors)

- I live in Stockholm on apartement . Jag är 29 . I live with my boyfriend . His name is Cezary . Apartement mine has a pattio and tree room . I enjoy there in Stockhulm but a lot of time to goto shop , fortifive minut . I have the buss and the Stockholm train . I lived in Danmark bifore , in Odense . It was less than Stockholm . I enjoy their too becaus I had more friends . I think it is hard to have friends here . But I enjoy better job here . In Odense jobbe I only on one website . In Stockholm I work on many website . I am webdevelooper . But Stockholm is closser to Luxembourg than Odense . It is important how one lives because I am not in my country . I mess my mother and my father but I live her with my boyfriend .

Future work

- Test on assistants → learn to incorporate automatically their corrections
- Integrate into Lärka and test on learners → collect their corrections into “reports”
- Adapt to English?
- Shared task?

Future work (for SVALA)

- Add (automatic) linguistic annotation and visualization
- Develop support for automatic error detection and labeling
- Make it available/modifiable for other languages/users (tagsets, versioning, server…)

GitHub repository

https://github.com/spraakbanken/swell-editor

A link to an article on SVALA

Wirén Mats, Arild Matsson, Dan Rosén, Elena Volodina. 2019. SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. *CLARIN-2018 post-conference volume*. LiUP Press.

http://www.ep.liu.se/ecp/159/023/ecp18159023.pdf

A link to an article on Anonymization

Beáta Megyesi et al. (2018). Learner Corpus Anonymization in the Age of GDPR: Insights from the Creation of a Learner Corpus of Swedish. *Proceedings of the 7th NLP4CALL workshop.* [pdf]

http://www.ep.liu.se/ecp/152/006/ecp18152006.pdf

A link to this presentation

http://tiny.cc/gximdz

Thank you!

- Questions?
- Comments?

undo redo

SVALA pseudonymization

show options

previous next

prev mod next mod

Filter / numeric label

1

Morphology
- gen
- def
- pl

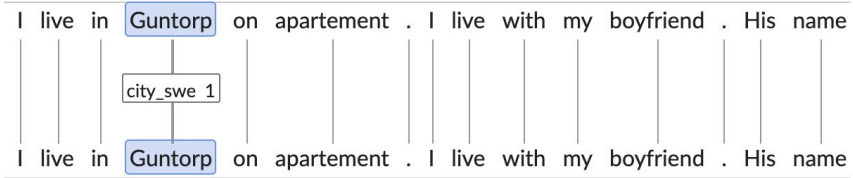
Names
- firstname_male
- firstname_female
- firstname_unknown
- surname
- middlename
- initials

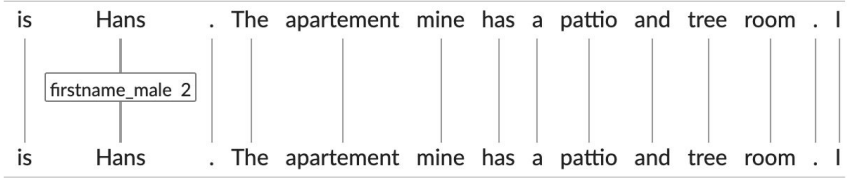
Geographic data
- country_of_origin
- country
- zip_code
- region
- **city_swe**
- city
- area
- place
- geo
- street_nr

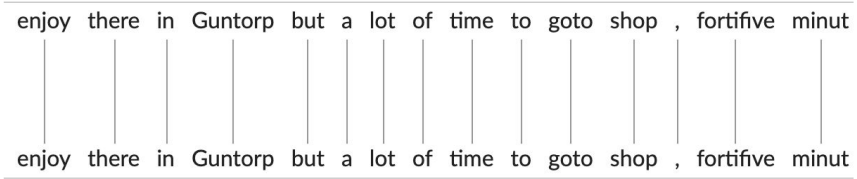
Institutions
- institution
- school
- work
- other_institution

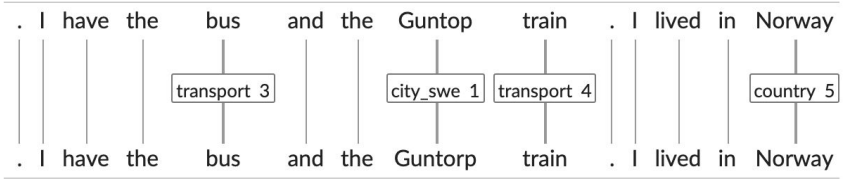
Transportation
- transport
- transport_line

Age
- age_digits









bifore , in Tromsö . It was less than Gunntorp . I enjoy their too

Document comment:

1
- Guntorp
- Guntop
- Gunntorp
- Guntorp
- Guntorp

2
- Hans

3
- bus

4
- train

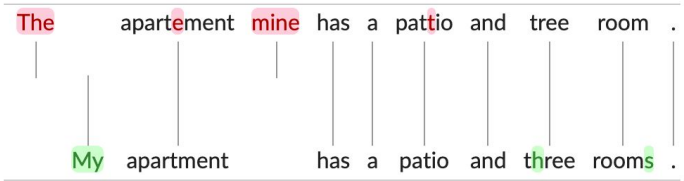
5
- Norway

6
- Tromsö

Target text:

I live in B-svensk-stad in an apartment . I live with my boyfriend . His name is Peter . My apartment has a patio and three rooms . I enjoy it there in B-svensk-stad but it takes a lot of time to go to the shop , forty-five minutes . I have the D-transport and the B-svensk-stad C-transport . I lived in C-land before , in D-stad . It was smaller than B-svensk-stad . I enjoyed it there too because I had more friends . I think it is hard to find friends here . But I enjoy a better job here . In D-stad I worked only on one website . In B-svensk-stad I work on many websites . I am a web developer . But B-svensk-stad is closer to B-land than D-stad . It is important how one lives because I am not in my country . I miss my mother and my father but I live here with my boyfriend .

Document comment:



undo redo

SVALA correction annotation

show options

previous next

prev mod next mod

group orphan

auto

Enter filter text

- L-M
- L-R
- L-REF
- L-W

Morphological

- M-ADJ/ADV
- M-CASE
- M-DEF
- M-F
- M-GEND
- M-NUM
- M-Other
- M-VERB

Punctuation

- P-M
- P-R
- P-W
- Sent-Segmentation

Syntactical

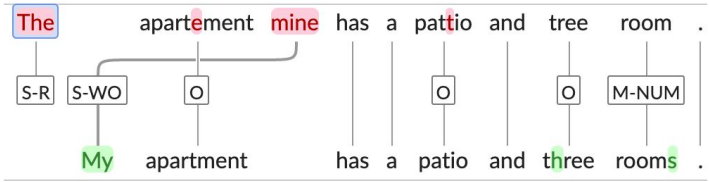
- S-adv
- S-CON
- S-finV
- S-M
- S-Msubj
- **S-R**
- S-W
- S-WO

Unidentified

- OBS!
- Uni

Target text:

I live in B-svensk-stad in an apartment . I live with my boyfriend . His name is Peter . My apartment has a patio and three rooms . I enjoy it there in B-svensk-stad but it takes a lot of time to go to the shop , forty-five minutes . I have the D-transport and the B-svensk-stad C-transport . I lived in C-land before , in D-stad . It was smaller than B-svensk-stad . I enjoyed it there too because I had more friends . I think it is hard to find friends here . But I enjoy a better job here . In D-stad I worked only on one website . In B-svensk-stad I work on many websites . I am a web developer . But B-svensk-stad is closer to B-land than D-stad . It is important how one lives because I am not in my country . I miss my mother and my father but I live here with my boyfriend .




```
{
  "source": [
    {"id": "s18", "text": "The "},
    {"id": "s19", "text": "apartement "},
    {"id": "s20", "text": "mine "},
    {"id": "s21", "text": "has "},
    {"id": "s22", "text": "a "},
    {"id": "s23", "text": "pattio "},
    {"id": "s24", "text": "and "},
    {"id": "s25", "text": "tree "},
    {"id": "s26", "text": "room "},
```


Document comment:

M-NUM

- room—rooms

O

- apartement—apartment
- pattio—patio
- tree—three

S-R

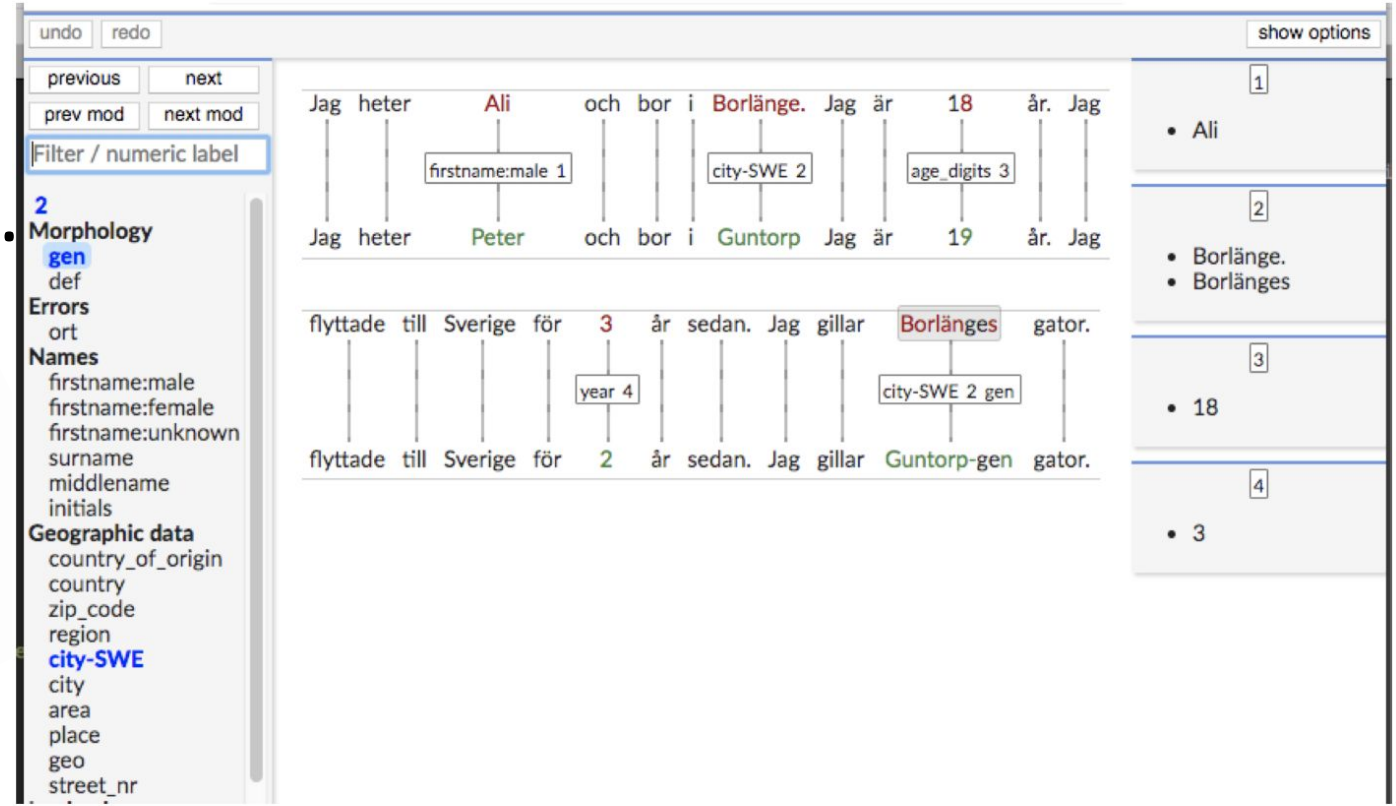
- The—

S-WO

- mine—My

SVALA pseudonym. mode

[Demo]



1. ORIGINAL TEXT → @PLACEHOLDER → RENDERING
2. ORIGINAL TEXT → @PLACEHOLDER → REPLACEMENT
3. ORIGINAL TEXT → @PLACEHOLDER → ORIGINAL

```
e-s6-t49": {
 "id": "e-s6-t49",
 "ids": ["s6", "t49"],
 "labels": ["city-SWE", "2"],
 "manual": true
,
```